1. Quasi-isometry inference with Batch Normalization

For batch normalization (BN) layer, its Jacobian, denoted as $J$, is not only related with components of activations (d components in total), but also with samples in one minibatch (size of $m$).

Let $x_j^{(k)}$ and $y_i^{(k)}$ be $k$th component of $j$th input sample and $i$th output sample respectively and given the independence between different components, $\frac{\partial y_i}{\partial x_j}$ is one of $m^2d$ nonzero entries of $J$. In fact, $J$ is a tensor but we can express it as a blocked matrix:

$$J = \begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1m} \\ D_{21} & D_{22} & \cdots & D_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ D_{m1} & D_{m2} & \cdots & D_{mm} \end{bmatrix}$$

(1)

where each $D_{ij}$ is a $d \times d$ diagonal matrix:

$$D_{ij} = \begin{bmatrix} \frac{\partial y_i^{(1)}}{\partial x_j} \\ \frac{\partial y_i^{(2)}}{\partial x_j} \\ \vdots \\ \frac{\partial y_i^{(d)}}{\partial x_j} \end{bmatrix}$$

(2)

Since BN is a component-wise rather than sample-wise transformation, we prefer to analyse a variant of Eq. 1 instead of $D_{ij}$. Note that by elementary matrix transformation, the $m^2d \times d$ matrices can be converted into $d m \times m$ matrices:

$$J = \begin{bmatrix} J_{11} & 0 & \cdots & 0 \\ 0 & J_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{dd} \end{bmatrix}$$

(3)

and the entries of each $J_{kk}$ is

$$\frac{\partial y_i}{\partial x_i} = \rho \left[ \Delta(i = j) - \frac{1 + \hat{x}_i \hat{x}_j}{m} \right]$$

(4)

The notations of $\rho$, $\Delta(\cdot)$ and $\hat{x}_k$ have been explained in our main paper and here we omit the component index $k$ for clarity. Base on the observation of Eq. 4, we separate the numerator of latter part and denote it as $U_{ij} = 1 + \hat{x}_i \hat{x}_j$.

Let $x = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_m)^T$, $e = (1, 1, ..1)^T$, we have

$$U = ee^T + \hat{x}x^T$$

(5)

and

$$J_{kk} = \rho(I - \frac{1}{m}U)$$

(6)

Recall that for any column vector $v$, $\text{rank}(vv^T) = 1$. According to the subadditivity of matrix rank [11], it implies that

$$\text{rank}(U) = \text{rank}(ee^T + \hat{x}x^T) \leq \text{rank}(ee^T) + \text{rank}(xx^T) = 2$$

(7)

Eq. 7 tells us that $U$ actually only has two nonzero eigenvalues, say $\lambda_1$ and $\lambda_2$, and we can formulate $U$ as follow:

$$U = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_1 \end{bmatrix}$$

(8)

combined with Eq. 6, finally we get the equation of $J_{kk}$ from the eigenvalue decomposition view, which is

$$J = P^T \rho \begin{bmatrix} 1 - \frac{\lambda_1}{m} & 1 - \frac{\lambda_2}{m} \\ 1 & \ddots \\ \vdots & \ddots & 1 \end{bmatrix} P$$

(9)
To show that $J_{kk}$ probably is not full rank, we formulate the relationship between $U^2$ and $U$

\[ U^2 = (ee^T + \hat{x}\hat{x}^T)(ee^T + \hat{x}\hat{x}^T) = ee^T ee^T + ee^T \hat{x}\hat{x}^T + \hat{x}\hat{x}^T ee^T + \hat{x}\hat{x}^T \hat{x}\hat{x}^T = mee^T + \left( \sum_{i=1}^{m} \hat{x}_i \right) eex^T + \left( \sum_{i=1}^{m} \hat{x}_i^2 \right) \hat{x}\hat{x}^T \]

\[ = mU + \left( \sum_{i=1}^{m} \hat{x}_i \right) eex^T + \left( \sum_{i=1}^{m} \hat{x}_i \right) eex^T + \left( \sum_{i=1}^{m} \hat{x}_i^2 - m \right) \hat{x}\hat{x}^T \]

(10)

Note that $\hat{x}_i \sim N(0, 1)$, so we can regard the one-order and second-order accumulated items in Eq. (10) as approximately equaling the corresponding one-order and second-order statistical moments for relatively large mini-batch, from which we get $U^2 \approx mU$.

The relationship implies that $\lambda_1^2 \approx m\lambda_1$ and $\lambda_2^2 \approx m\lambda_2$. Since $\lambda_1$ and $\lambda_2$ cannot be zeros, it concludes that $\lambda_1 \approx \lambda_2 \approx m$ therefor $1 - \frac{\lambda_1}{m} \approx 0$ and $1 - \frac{\lambda_2}{m} \approx 0$ if batch size is sufficient in a statistical sense.

References